Case Studies 2022L

Residual-diagnostics plots

May 5, 2022

Residuals

For a continuous dependent variable Y, residual r_i for the i-th observation in the dataset is the difference between the observed value of Y and the corresponding model prediction:

$$r_i = y_i - f(\underline{x}_i) = y_i - \hat{y}_i$$

Standardized residuals are defined as

$$\tilde{r}_i = \frac{r_i}{\sqrt{Var(r_i)}}$$

where $Var(r_i)$ is the variance of the residuals r_i

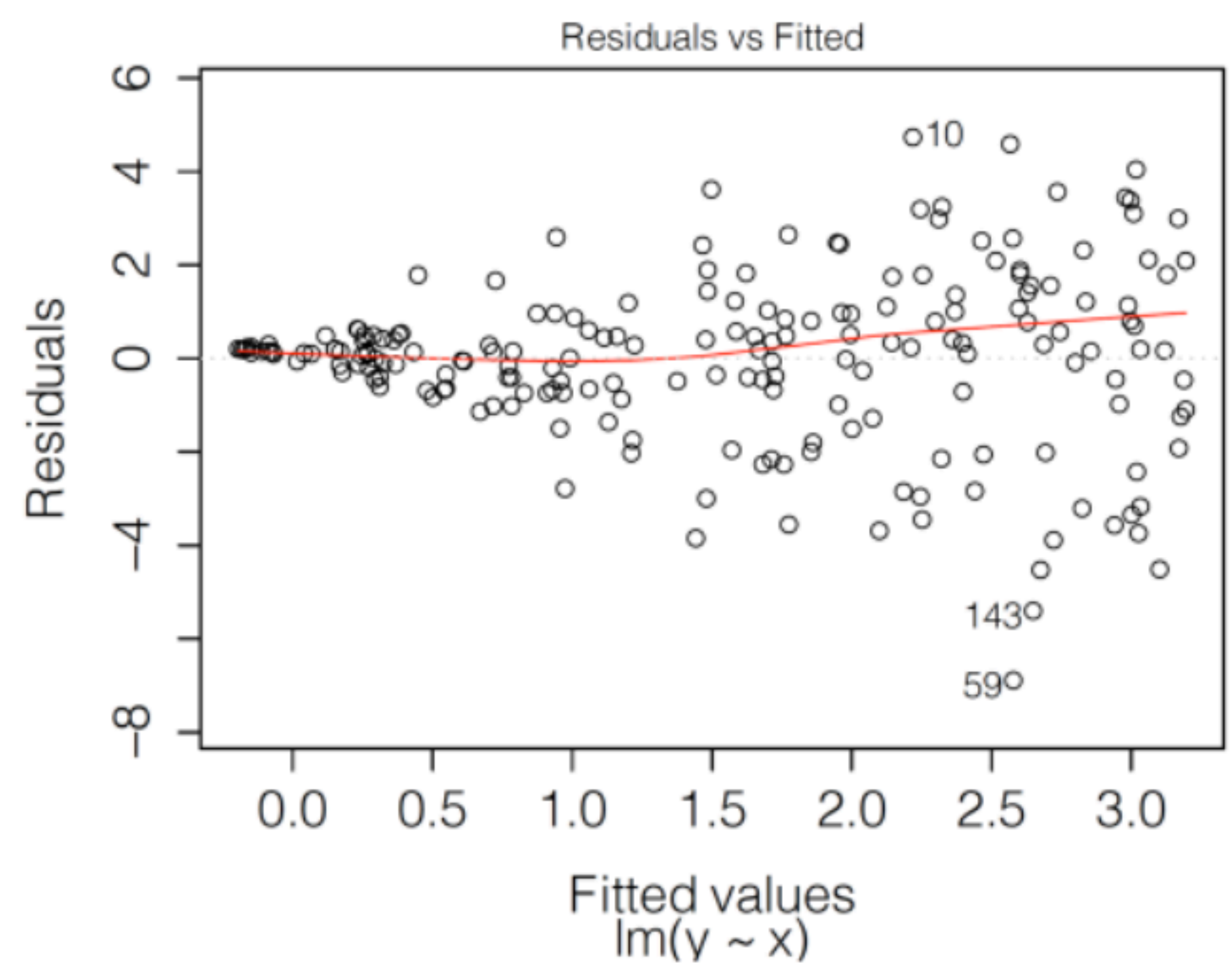
Introduction

Residuals can be used to identify potentially problematic instances. The single-instance explainers can then be used in the problematic cases to understand, for instance, which factors contribute most to the errors in prediction.

Residuals should express a random behavior with certain properties. If we find any systematic deviations from the expected behavior, they may signal an issue with a model.

Residuals-fitted values plot

- Plot presents the residuals in function of the predicted (fitted) values.
- For a well-fitting model, the plot should show points scattered symmetrically around the horizontal straight line at 0.
- However, the scatter in the plot has got a shape of a funnel, reflecting increasing variability of residuals for increasing fitted values.
- This indicates a violation of the homoscedasticity, i.e., the constancy of variance, assumption.
- Also, the smoothed line suggests that the mean of residuals becomes increasingly positive for increasing fitted values.
- This indicates a violation of the assumption that residuals have got zero-mean.

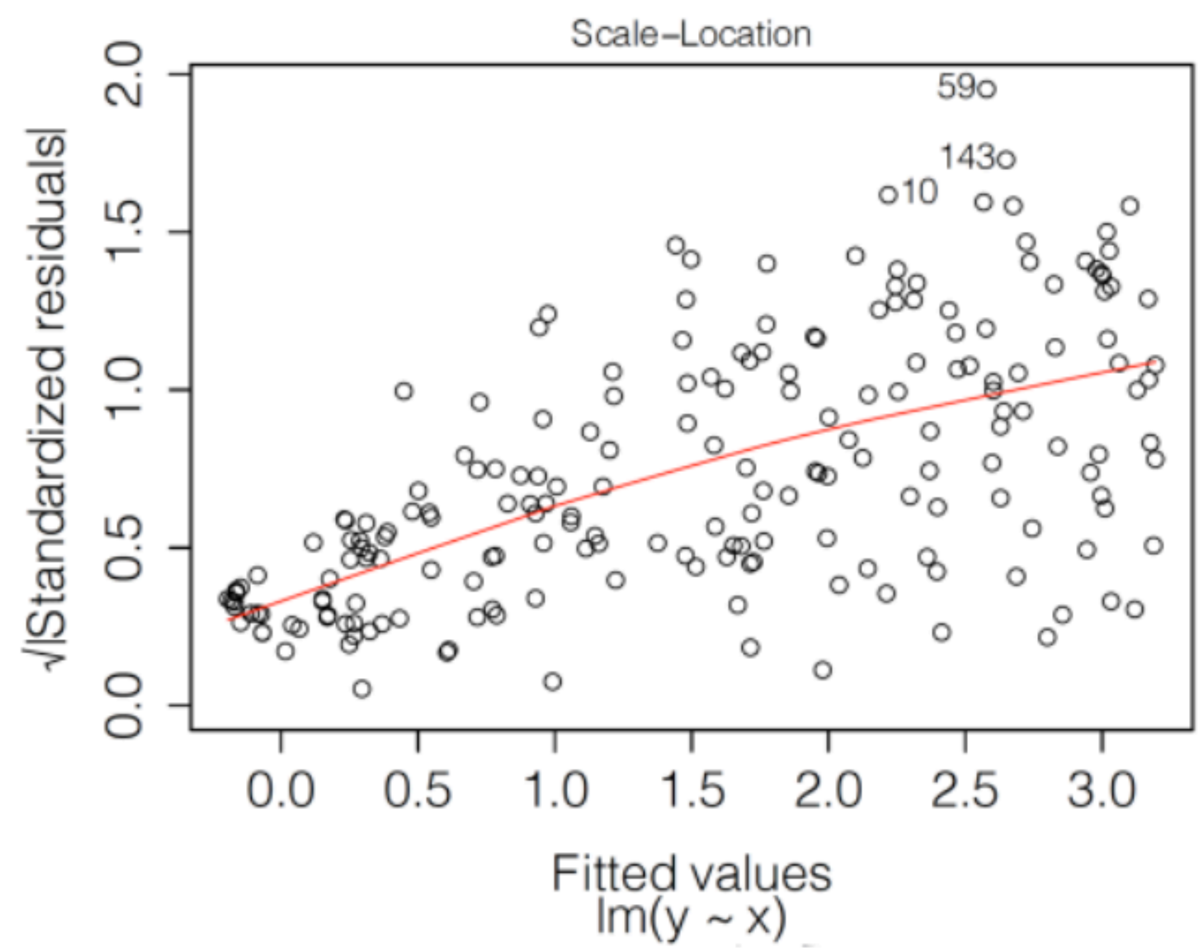


Scale-location plot

The scale-location plot presents the relationship between squared residuals and fitted values.

For a well-fitting model, the plot should show points scattered symmetrically across the horizontal axis.

This is clearly not the case of the plot, which indicates a violation of the homoscedasticity assumption.

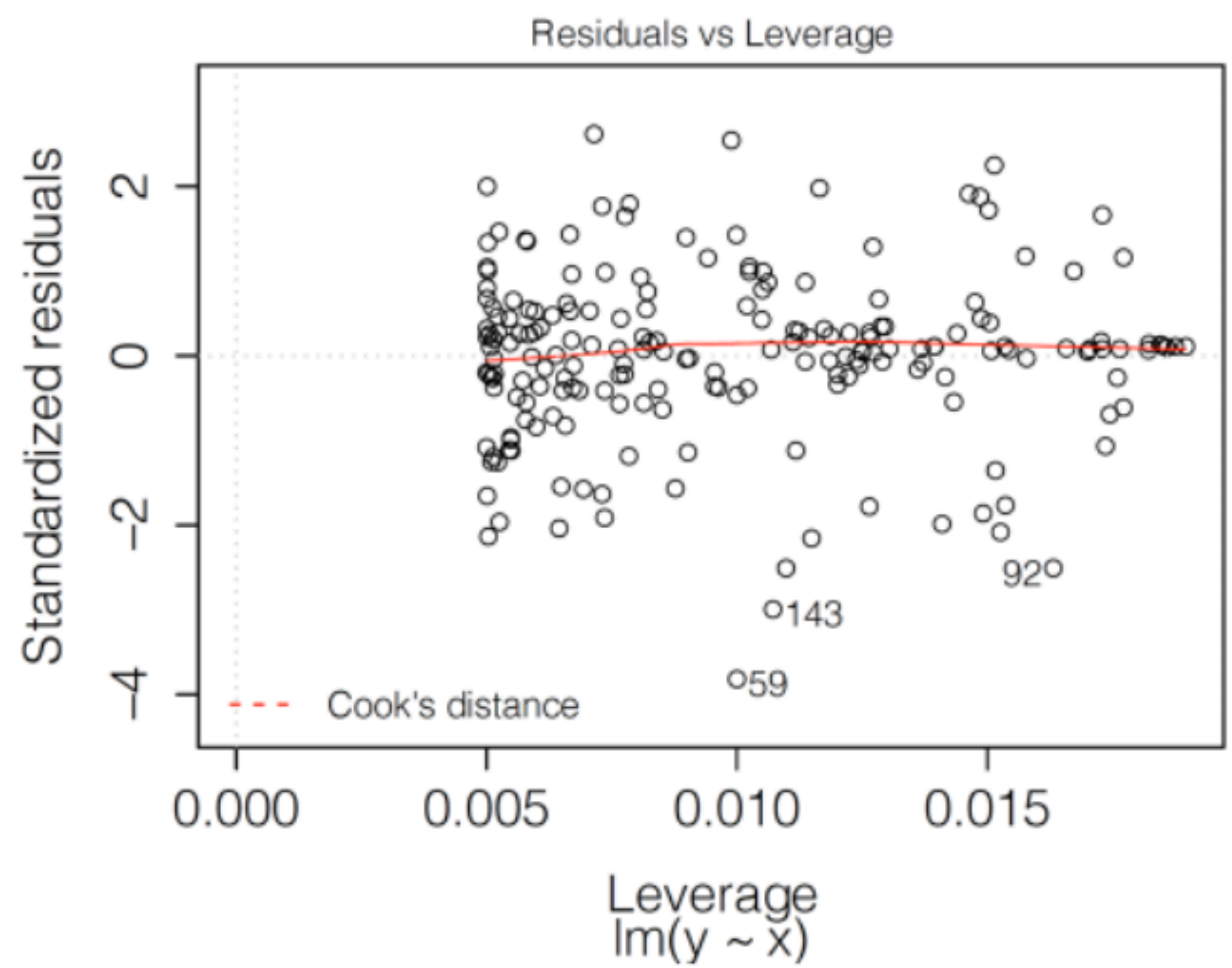


Residual-Leverage Plot

Plot presents the relationship between standardized residuals and *leverage.*

Leverage is a measure of the distance between explanatory variables and the vector of mean values of all explanatory variables (Kutner et al. 2005).

A large leverage value for the i-th observation is distant from the center of all observed values of the vector of explanatory variables. A large leverage value implies that the observation may have an important influence on predicted/fitted values.

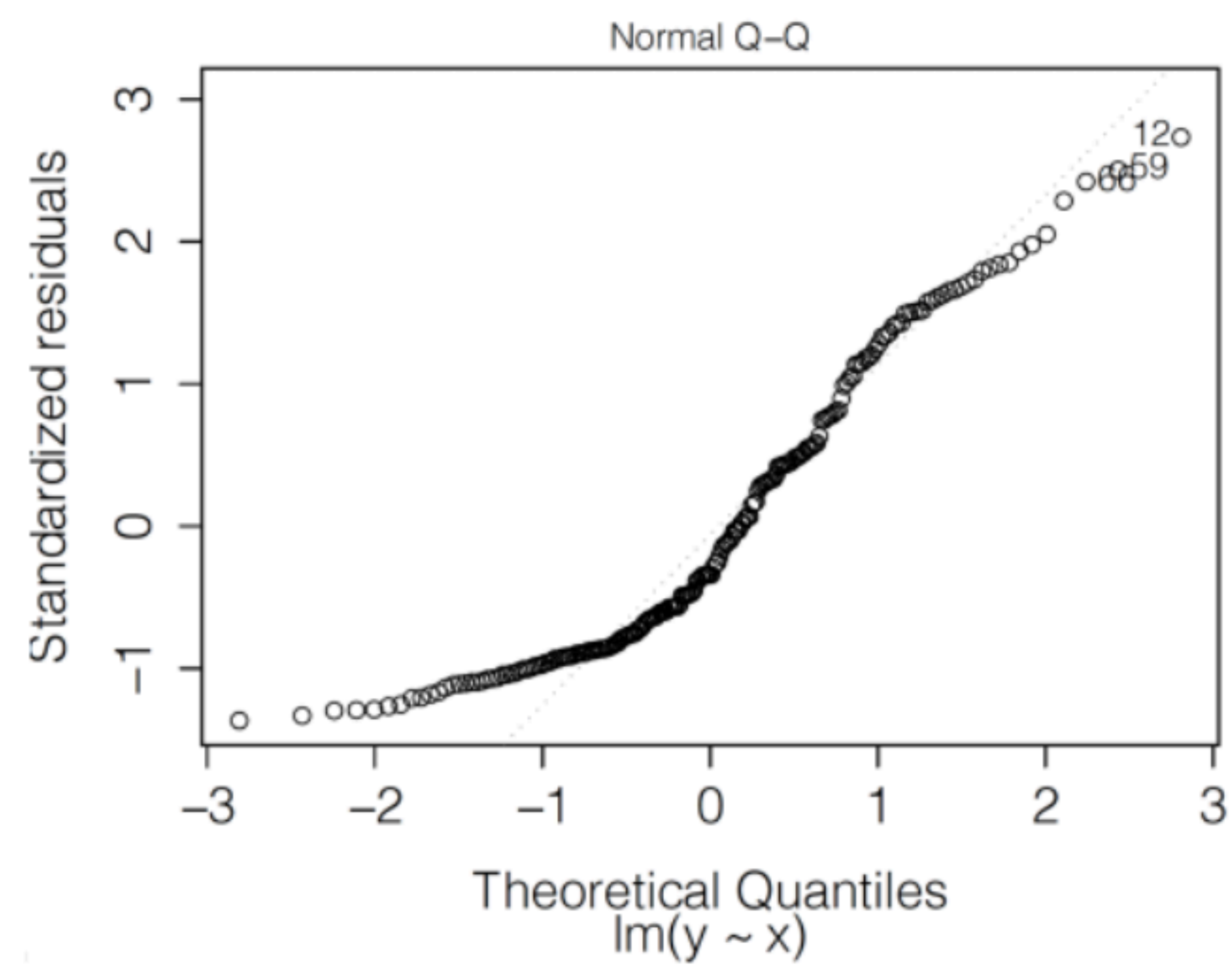


Normal Q-Q Plot

- The vertical axis represents the ordered values of the standardized residuals, whereas the horizontal axis represents the corresponding values expected from the standard normal distribution.

- If the normality assumption is fulfilled, the plot should show a scatter of points close to the y=x line.

- Clearly, this is not the case of the plot.



Pros and Cons

Diagnostic methods based on residuals are very useful tool in model exploration to identify different types of issues with model fit or prediction:

- Distributional assumptions
- Detecting groups of observations for which a model's predictions are biased and require inspection.

Pros and Cons

A potential complication related to the use of residual diagnostics is that they rely on graphical displays:

- One may have to construct and review many graphs for a proper evaluation of a model.
- Interpretation of the patterns seen in graphs may not be straightforward.
- It may not be immediately obvious which element of the model may have to be changed to remove the potential issue with the model fit or predictions.

References

Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li. 2005. *Applied Linear Statistical Models*. New York: McGraw-Hill/Irwin.

Gosiewska, Alicja, and Przemyslaw Biecek. 2018. auditor: Model Audit - Verification, Validation, and Error Analysis. https://CRAN.R-project.org/package=auditor.

Please feel free to send e-mail about your questions!

📩 mustafa.cavus@pw.edu.pl